Spontaneous speech in the teaching of phonetics and speech perception

Natasha Warner[1], Seongjin Park[2]

[1,2]University of Arizona, U.S.A,
[1]*nwarner@email.arizona.edu*, [2]*seongjinpark@email.arizona.edu*

Abstract

Speech contains vast variability, more than we often expect based on phonetics courses. Spontaneous, conversational speech in particular is frequently realized with deletions and alterations to many of the expected sounds. We refer to this as reduced speech.

Reduced speech in context is quite perceptible to native listeners, but less so to non-native learners. Since L2 learners will encounter reduced speech when they leave the language classroom, spontaneous speech presents an opportunity for teaching perception of more natural speech to L2 learners. Furthermore, spontaneous speech can be quite intriguing for phonetics and linguistics students, so it also presents an opportunity for engaging students at all levels with language and phonetics.

Keywords: Reduced speech, spontaneous speech, perception, teaching of phonetics.

1. Introduction

Spontaneous speech does not contain a neat series of sounds that look like the examples we typically teach in phonetics courses. We may teach students in phonetics courses that a voiceless stop has a silent closure followed by a burst and possibly aspiration noise before voicing restarts, or we may teach that a voiced stop has a period of low-amplitude voiced closure with a gap in the formant structure. However, when one examines recordings of spontaneous speech in detail, one can readily find examples where a phonemically voiceless or voiced stop is not realized this way at all, but is instead realized as an approximant with formants continuing throughout the consonant and no burst, as well as tokens where the consonant is deleted, possibly leaving some effect on neighboring sounds (Fig. 1). We refer to such speech, where segments are often realized with a different manner of articulation than usual and some segments and syllables are deleted, as reduced speech. Since reduced speech is not the exception but the norm, and is the common material of daily life communication, it is important to understand what realizations speakers are actually producing in their normal speech. We find that spontaneous speech, which frequently contains reduced speech, is useful for teaching linguistic phonetics, and also presents an opportunity and a challenge for teaching speech perception in the language classroom.

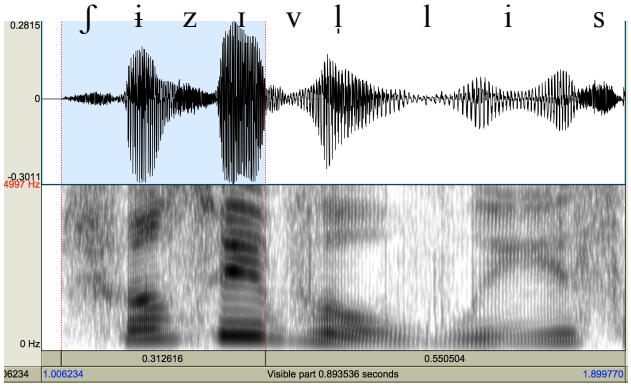

Figure 1: Spontaneous conversation recording of the words "she wants to be a police" (followed by "officer, I think"), in which "she wants to be a" (selected portion) is highly reduced and the /p/ of "police" is realized as a voiced fricative. Audio of this and other examples available at https://nwarner.faculty.arizona.edu/content/6

Reductions like this happen in every language where spontaneous speech has been examined for them, as far as we are aware [1-4]. Figure 2 demonstrates a Japanese example, where the number of moras is altered and one stop is (nearly) deleted.

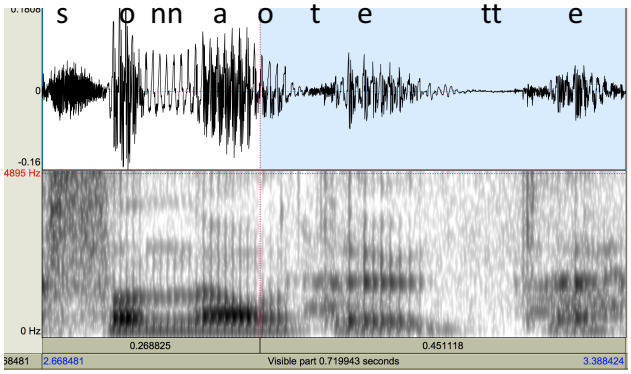

Figure 2: Spontaneous conversation recording of the Japanese words /sonna koto itte/ (また)そんなこと言って 'and say something like that (again)', in which /k/ of /koto/ is effectively deleted and /oi/ is merged to [e].

Reductions such as these are pervasive in spontaneous speech, and reductions of a single consonant are not rare even in more careful speech

10.21437/ISAPh.2018-6

such as news interviews or formal meetings. On examining the waveform and spectrogram of a casual, spontaneous conversation, it usually does not take long to identify highly reduced portions in most speaker's speech. This raises questions both about how we teach students linguistic phonetics, and about how to teach language students oral comprehension. It also provides a highly engaging way to draw students in to the study of speech and linguistics.

2. Spontaneous speech in teaching of linguistic phonetics and general linguistics

2.1. Spontaneous speech for engaging interest

Spontaneous speech containing severely reduced portions is highly engaging for phonetics classes, introduction to linguistics classes, and outreach talks that bring speech and linguistics to the general community. To prepare for this activity, either record a spontaneous conversation between two friends who are native speakers of the language most of the students speak natively, or find a corpus of casual spontaneous speech in that language. Listen to portions of the recording in Praat [5] or other speech analysis software, making sure to zoom in and examine what phonetic segments were actually produced. Even phoneticians, on listening to a whole utterance, tend to recognize the words in context and not to realize how much reduction was produced.

Once one has identified a few utterances containing massive reductions [6, 7], one should identify the most reduced words. Listen to the whole utterance a few times to be sure of what the words are. Then select various portions of the utterance and play them, in order to determine what sounds and words are contained in the clearer portions of the speech. Whatever unclear portions of speech are left must represent the remaining words/sounds.

For example, in Figure 3, one might select the portion before the highlighted portion and recognize the word /so:/, and select portions after the highlighted portion and recognize /jo: no/ and /supe:su/. The selected portion itself is difficult to recognize, but it must then be whatever realization was produced for /ju: hitotati/. If one selects the portion before the highlighted portion plus the remainder of the voicing before the voiceless fricative, one hears something like [soiç] or [soiɟ]. One might not recognize this as /so: ju:/, although it is a legitimate realization of /so: ju:/ in the context of a following [ç] (/h/), so this method requires selecting and listening to a wide variety of portions with more and less context. Noting here that [çəʃ] is a highly unexpected realization for /hitotati/ (realized as [çitotatɕi] in careful speech), one might select just /hitotati/ or /ju: hitotati/ (the highlighted portion of the figure) and save this as a separate file. For English examples, the first author has several on her website which readers are welcome to use in teaching (https://nwarner.faculty.arizona.edu/content/6).

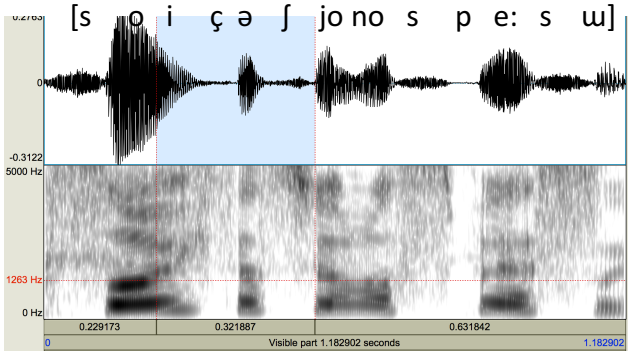

Figure 3: Spontaneous conversation recording of the Japanese words /so: ju: hitotati jo: no supe:su/ そういう人たち用のスペース 'Space for people like that,' with /hitotati/ realized approximately as [çəʃ].

In class, one can play first the extracted short portion (one word or a few words) and ask students what they think they heard. State what language it is, and perhaps how many words are included. After allowing students to guess what was said, play the larger context file. In most cases, students who are native speakers of the language will not be able to recognize the words correctly in the short file (out of context). However, they will be able to recognize which portion of the longer file corresponds to the short file and what words it is, as soon as they hear it in the context of the longer file. They will also be able to hear that the longer file is very natural casual conversation, and does not sound strange or disfluent. This tends to leave the students amazed, and quite engaged with the topics of speech, phonetics, speech perception, and spoken word recognition. After playing the highly reduced speech out of context and in context, all it takes to get students into these topics is to ask "So how do we understand that?"

The authors can offer a few tips on using reduced speech demos in class. First, when saving the short and long files, make sure to use a filename that does not contain the words in question, so that students do not see the answer displayed on the screen if one opens the files in Praat during class. Second, some speech reductions are so extreme that listeners cannot even locate the reduced portion they have just been listening to in the longer context when they hear it. Very rare tokens are so reduced that even native listeners do not recognize the words even in context. These examples do not work well to make the point about reduced speech being a normal way to convey information. Finally, it is best to choose a (partial) utterance that is not too long as the surrounding context, so that finding the reduced portion and remembering the words does not become confusing.

The first author has used this demonstration with many reduced speech samples with a wide variety of classes and even with community outreach events, as well as in academic talks. This demonstration is usually engaging for first-year university general education classes that have never studied any type of phonetics, for graduate level phonetics courses and advanced graduate seminars, and for classes in between such as Introduction to Linguistics. It is also very successful with outreach visits to high school classes and would probably also be successful with younger children. That is, this type of demonstration can be very useful for spreading awareness of linguistics and phonetics to the general public. However, it might be less useful at a science fair or festival setting, because of the high level of background noise at such events.

Above, we have described using reduced speech demonstrations with the speech being in the native language of most of the students, in order to bring up topics like phonetic variability, spoken word recognition, and transmission of information. A second use of such demonstrations is for teaching language learners what kinds of variability they might encounter in their L2. Most L2 learners in the classroom hear primarily careful, formal, non-native-listener-directed speech, often from their teachers. One might assume that they will primarily hear careful speech, directed to them, even when they leave the classroom and go out into the L2-speaking society, and that such careful speech will not contain reductions. However, even formal speech can contain reductions. We have noticed reductions of individual stops to approximants or deletions of stops in the speech of a news interviewer ([sɛ̃nəɚ] for 'senator') and the speech of a university president addressing a formal meeting ([lɪɚl̩i] for 'literally'), and we recorded a token of [kʌlɨn ɚpʰɔɪɾ ɨ nɛwɚk ɪʃʊ] for 'calling to report a network issue' from a university tech support answering machine message. It is likely that there are some tokens of massive reduction, deleting entire syllables, even in formal speech, although not as many as in casual conversation. Given this, non-native learners of English who study abroad at an English-speaking university are likely to encounter some tokens with substantial reduction, even if people often address them in non-native-listener-directed speech. Of course, as soon as they go out into the broader English-speaking society, they will encounter reduced speech frequently.

Therefore, for teaching languages in the classroom, it could be very useful to play demonstrations of highly reduced speech in the target L2 language. We have probably all had the experience of studying a language in the classroom for some years, and then going abroad to the country where the language is spoken, and having a rude awakening when we could not understand anything in a casual conversation between native speakers. While teaching L2 learners about examples of highly reduced speech may not train them to be able to understand it, it will at least make them aware of what casual natural speech in the language sounds like. It is probably also important to emphasize when using highly reduced speech demonstrations in the L2 classroom that such reduction is normal, and happens in the students' native language as well. Otherwise students may simply conclude "English is very strange and impossible to learn" or "English speakers have sloppy pronunciation" instead of realizing the scope of phonetic variability in any language. When the first author has played samples of reduced speech to groups that include non-native listeners, the non-natives often find it difficult to understand the speech even in utterance context.

2.2. Teaching transcription

If teaching a phonetics or introduction to linguistics course that includes transcription practice, transcribing spontaneous speech provides an intriguing challenge for students. Having to determine a narrow phonetic transcription of highly reduced speech may prompt students to realize just how reduced it is. However, students are often still influenced by how they expect a lexical item to be pronounced.

The first author has assigned students in an introductory graduate level phonetics course to transcribe spontaneous speech, such as the utterance "I can't register in person, so they're just gonna hafta deal with that" (available at https://nwarner.faculty.arizona.edu/content/6). The first author's own transcription of the "they're just" portion is [ðe̞i̯ sː], and the author's transcription of the "gonna hafta" portion as recorded is [kɯɻ̟ æ̟ $t^{(h)}$i̯̟]. Parentheses represent a symbol that there is only weak evidence for. Students had full access to the sound file and could examine the spectrogram and play any portions they found useful.

Although some students realized there was massive reduction somewhere in this sequence, many still transcribed a vowel in the word "just" that was not present. An anecdotal sample of the transcriptions given by four native English speaking students includes [ðɛr dʒɪs, ðə dʒəs, ðɛɪz s, leɪ ʒs]. All four included more distinct types of frication than the author identified, and two added a vowel nucleus. These students transcribed the "gonna hafta" string as [gʌ̃ʰ tʰɪ, gʌ̃na tə, gn̩ ɻ̟̊f tʰə, gə̃ æftə]. Three of these four transcribed three syllables, where the spectrogram and perceptual cues support just two. Two of these transcriptions include an [f], which is

not supported by the spectrogram. There is no one single correct transcription for sequences like these, and all of these students correctly detected considerable reduction, but their transcriptions do seem to be influenced by lexical expectations of how "gonna hafta" would be pronounced in careful speech, i.e. [gʌ̃nə hæftə]. This suggests that even graduate-level phonetics students with full access to the sound file, who have been warned about reduced speech, are unable to ignore lexical expectations. The influence of what sounds one expects to hear is quite strong.

Depending on the topics of the class, one could extend the pedagogical uses of spontaneous speech to various other topics. One could use a sample of highly reduced speech to launch a discussion of spoken word recognition, since it is not clear how listeners can recognize either 'going to have to' or 'gonna hafta' (assuming that is lexicalized) from [kɥæ̰ $t^{(h)}$i̯] in most models of spoken word recognition. One could launch a discussion of what kinds of information listeners may use to recognize words, and of whether speech is "sloppy" or "efficient" as a means of communication. One could also ask students how highly reduced speech affects non-native listeners. This ties in to the topic below.

3. Effect of spontaneous speech on L2 listeners: Transcription task

Given that even native listeners cannot recognize highly reduced speech out of context, how does the reduction common to spontaneous speech affect non-native listeners? We have probably all had the experience of listening to a fast, casual conversation in our non-native languages and feeling lost. We might attribute this to fast speech rate or unfamiliar vocabulary, but such speech is also likely to contain massive reductions that we may fail to understand in an L2. In a larger project that this work forms a part of, we examined how well native and non-native listeners, and native listeners of a different dialect, can perceive and recognize the words of spontaneous American English conversation.

3.1. Methods

The detailed methods for this work will be described in future publications of the research group, and are too lengthy to cover fully here. Briefly, 30-95 listeners from various language groups listened to utterances extracted from spontaneous, casual American English conversations. The speakers in the conversations were talking on the phone, with one speaker (the only one recorded) seated in a sound protected booth and recorded through a high-quality head-mounted microphone. The speakers in each conversation were friends or family members, so the utterances were very casual. One utterance from each of 13 speakers, each containing substantial speech reduction, was chosen. The methods for recording these conversations are described in Warner & Tucker [8]. Each listener heard the utterances over headphones, while seated in a sound protected booth, and typed into a computer what words they heard. They then heard the utterance a second time, and had a chance to correct their typed response.

Each response was evaluated for the percentage of words in the utterance that failed to appear in the response (Word Error Rate, WER). An additional measure of Edit Distance accounted for errors other than leaving out target words, such as reversing the order of words or inserting words. However, these two measures were highly correlated, so either is representative of the results.

The listener groups were native speakers of: American English (students at University of Arizona, most from the Southwest or California), Canadian English (students at University of Alberta), New Zealand English (students at University of Canterbury), Dutch, Japanese, and Spanish. All but the Japanese and some Spanish groups were living in their home country at the time of the experiment (e.g. the Dutch in the Netherlands, the New Zealand English speakers in New Zealand). The Japanese speakers were living in the U.S. or Canada. The Spanish speakers constituted three groups: 1) Early Spanish-English bilinguals who were born in the U.S. or immigrated by age 5, who mostly grew up in Southern Arizona, who were students at the University of Arizona, 2) Late Spanish-English bilinguals who immigrated to Arizona after age 17 (mostly non-students), and 3) Students at the University of Querétaro in Mexico, who were living in Mexico at the time of the experiment. The former two groups were living in Arizona at the time of the experiment and using Spanish to varying degrees in their daily lives.

3.2. Results

Figure 4 presents the overall results demonstrating how well each listener group was able to transcribe the words of the conversational speech stimuli.

Figure 4 demonstrates that all of the groups with native competence in English, regardless of variety of English and regardless of whether they also have native competence in Spanish, transcribed the words of conversation rather accurately, omitting or replacing approximately 10% of the words the speakers produced. The native Dutch listeners show a higher error rate, exceeding 20% of the words incorrect. The remaining three groups (Spanish-speaking late immigrants, Spanish-speaking students living in Mexico, and Japanese native speakers) all

score poorly, with the Japanese listeners missing approximately 60% of all words uttered.

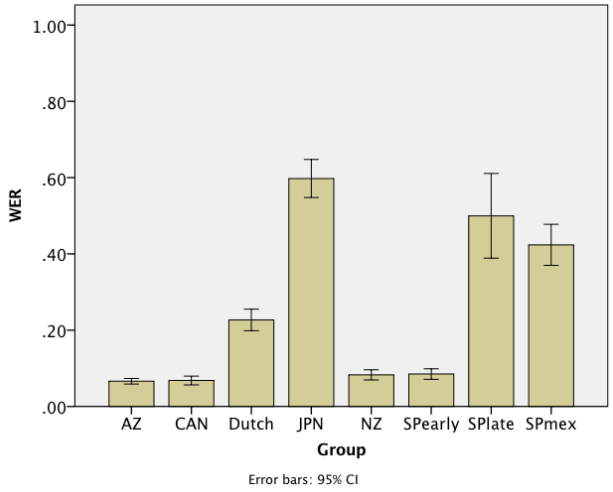

Figure 4: Average Word Error Rate (WER) for each listener group, across the 13 conversational speech stimuli for the transcription task. "AZ" refers to monolingual native speakers of American English living in Arizona. "SPearly" refers to the Spanish-English bilinguals who were born in or immigrated early to the U.S., "SPlate" to the group of late immigrants, and "SPmex" to the students at a university in Mexico.

We will examine some examples of listeners' responses qualitatively, to investigate what listeners are actually recognizing when they hear spontaneous, casual speech, especially in their non-native language. In an actual conversation between two people, the listeners would be likely to understand more words than they are able to transcribe correctly in our task, since they would know all the previous context of the conversational topic. However, in real life situations such as joining friends' conversation mid-way through or overhearing speech directed to someone else, the experimental task may be quite representative of how well non-native listeners can recognize words in spontaneous speech.

First, we should acknowledge that even native listeners do not always perceive (or at least report) all words of an utterance correctly. For the stimulus "Now she's deciding we could probably wear like black and white" one native listener of the Arizona group responded with "Now she's deciding and if we're probably black and white," with an error rate of 36% and a rather altered meaning. The listener seems to have misperceived "wear" as "we're," altering the sentence structure. For the stimulus "we go to lunch, or go to the park with the dogs, and" one listener of this group reported "We go to l and church and look out with the dogs" (WER: 46%). However, such poor perceptions by native listeners are relatively rare.

Sometimes listeners understand the meaning of a stimulus but do not accurately report the words. For the stimulus "You don't understand how confusing this building is. It's like a maze" one native listener of the Arizona group responded "You don't know how confusing this building is, it's like a maze," (WER: 8%) apparently misremembering "understand" as "know." This reflects memory limitations in this task, rather than speech perception. However, the measure of WER does give an approximate reflection of both speech perception and understanding.

Non-native listeners sometimes find it difficult to report very many words of a stimulus at all. For example, for the stimulus "Cause we're probably gonna hafta have somebody come inspect it, y'know?" one Japanese listener responded with just "Cause I probably" (WER: 82%). For the stimulus above about going to lunch, one Japanese listener responded with "we got a rain" (WER: 92%, only "we" is correct). Sometimes non-native listeners failed to give any response at all, even though they were given two chances to hear the entire utterance.

Sometimes non-native listeners succeed in reporting a larger number of words, but with a wide range of alterations to the intended word string. For the stimulus about going to lunch, one Spanish speaker (Mexico group) responded with "record of the line to the park of the darkside" (WER: 69%). For the stimulus "It makes me look like a mom, and I don't want anything that makes me look like a mom" a Japanese listener responded with "and it makes me quite mad, and it'll make me quite mode" (WER: 74%). For the stimulus above about having someone come inspect something, one Japanese listener responded "cause it has come that somebody comes in the other, you know" (WER: 50%). A Dutch listener responded to the same item with "so if any one comes and inspect it you know" (WER: 67%).

However, listeners from the same groups sometimes perform rather well. One Japanese listener responded to the stimulus above about a confusing building with "You don't understand how confusing spelling is. It's like amazing" (WER: 33%). For the stimulus "We were like yeah, that's probably not a good idea anymore" one Japanese listener responded with "Or like yeah probably that's not a good idea any more" (WER: 17%). For the stimulus about having someone inspect something, one Spanish Late Immigrant responded "Cuz we're probably going to have somebody come inspect it you know" (WER: 8%). All listener groups do have some listeners who score 0% error on some stimuli.

Some errors are explicable through phonetic or lexical facts. The confusion of "this building" with "spelling" or "the spelling" above is one that many listeners of various language groups made, and it is phonetically motivated by the fact that English word-initial /b/ and post-s /p/ are both voiceless unaspirated, and by reduction of the /d/ in "building" in this stimulus. The confusion of "a maze" as

"amazing" is also common across many listener groups, and probably reflects word frequency and perhaps unfamiliarity of the word "maze." One could perhaps analyze a phonemic transcription of the words in the response in comparison to a phonemic or phonetic transcription of the segments actually produced in the stimulus to get a measure of phonological rather than word-based similarity of stimulus and response. However, explanations of what might have motivated a particular misperception are always bound to involve some speculation.

Our research group has begun investigating what sociolinguistic factors might correlate with WER within various language groups and across all groups. However, it appears to be difficult to ask sociolinguistic background questions that truly give comparable answers across all language groups, including both immigrants and students living in their own country, across a variety of cultures. For example, we asked all bilingual and L2 English groups both at what age they began learning English and at what age they first felt comfortable speaking English. The Japanese and Dutch listeners appear to have answered the former question based on when they started taking English in school, with almost all listeners stating the same age based on their country's educational system. The Spanish students at a Mexican university and the Spanish late immigrants gave far more varied answers, ranging almost up to their current age, even though at least the Mexican university students presumably also started learning English in school at a relatively consistent age. For the question about when they became comfortable speaking English, approximately half of all Japanese listeners responded that they were not yet comfortable in English, even though they were living in an English-speaking country at the time of the experiment and functioning in English. No other group responded this way so often. Our initial conclusion from a basic analysis of correlations between WER and sociolinguistic factors is that although we attempted to devise questions that would work across languages and countries, the differences in educational system and sociolinguistic situation make these results difficult to compare across groups. However, further analysis will likely reveal some patterns. Because our current results on this are tentative, we leave this for future research.

4. Conclusions

Several patterns emerge from the perception data. We see that both native and non-native listeners vary widely in how accurately they can perceive and report the words of a spontaneous speech utterance. All groups sometimes show perfect perception of a stimulus by some listener, and all groups (including natives) sometimes show quite poor perception of a stimulus. It is also clear that one does not have to correctly perceive every word of an utterance in order to understand the content. However, both native and non-native listeners also often alter the meaning of utterances through their reporting. In future research, we plan to investigate how accurately listeners' responses recreated the original meaning.

Turning to the connection between spontaneous speech reduction and second language learning and teaching, this sort of transcription data can show us how much or how little language learners are understanding once they leave the carefully controlled L2 classroom and go out into the larger L2-speaking society. They are probably missing a lot of the words, especially when they hear spontaneous speech that is not addressed specifically to them. However, as noted above, one does not have to correctly perceive every word to understand speech. Still, seeing the degree of misunderstanding of words could lead teachers in language classrooms to incorporate more practice with perceiving spontaneous speech into their teaching.

Spontaneous speech also presents many opportunities for teaching in the linguistic phonetics classroom, as well as for community outreach, teaching general education introductory courses, and teaching phonetics and psycholinguistics at advanced levels. The same reductions that pose problems for non-native learners can be especially effective for getting audiences (students or otherwise) engaged with phonetics and linguistics. Phonetics and psycholinguistics as fields have many intriguing demonstrations that work well with audiences, such as the McGurk Effect and the pie-buy-spy demonstration of VOT perceptual cues. Using spontaneous speech as a demonstration is likely to work for any language, and it relates to listeners' daily life experience of speech. Thus, it adds a useful tool to our inventory for teaching.

5. Acknowledgements

We would like to thank Mirjam Ernestus, Benjamin Tucker, Miquel Simonet, and Dan Brenner for helpful discussion of these topics and for their help in collecting parts of the data. We would also like to thank the many other people who facilitated the data collection in various ways. We thank the attendees of the ISAPh conference for their helpful discussion of this work. This work was funded by NSF grant #1022313 to the first author and Co-PI Simonet.

6. References

[1] *Barry, W., Andreeva, B.* 2001. Cross-language similarities and differences in spontaneous speech patterns. *J. Int'l Phonetic Assoc*. 31, 51-66.

[2] Arai, T. 1999. A case study of spontaneous speech in Japanese. *Proc. 14th ICPhS* Berkeley CA, 615-618.
[3] Ernestus, M., Warner, N. 2011. An introduction to reduced pronunciation variants. *J. Phonetics*, *39*(SI), 253-260.
[4] Kohler, K.J. 1999. Articulatory prosodies in German reduced speech. *Proc. 14th ICPhS* Berkeley CA, 89-92.
[5] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341-345.
[6] Johnson, K. 2004. Massive reduction in conversational American English. In: K. Yoneyama, K. Maekawa (eds), *Spontaneous speech: Data and analysis. Proc. 1st Session, 10th Int't Symp.* Tokyo: Nat'l Int'l Inst. Japanese Language, 29-54.
[7] Greenberg, S. 1999. Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Comm.* 29, 159-176.
[8] Warner, N., Tucker, B.V. 2011. Phonetic variability of stops and flaps in spontaneous and careful speech. *J. Acoust. Soc. Am.,* 130, 1606-1617.